International Journal of Epidemiology, 2015, 117–128
doi: 10.1093/ije/dyu215
Advance Access Publication Date: 30 October 2014
Original article

Cardiovascular Disease and Cardiovascular Risk Factors

Subgroups at high risk for ischaemic heart disease: identification and validation in 67 000 individuals from the general population

Ruth Frikke-Schmidt,[1,2]* Anne Tybjærg-Hansen,[1,2,3] Greg Dyson,[4] Christiane L Haase,[1] Marianne Benn,[5] Børge G Nordestgaard[2,3,6] and Charles F Sing[7]

[1]Department of Clinical Biochemistry, Rigshospitalet, [2]The Copenhagen General Population Study, Herlev Hospital, [3]The Copenhagen City Heart Study, Frederiksberg Hospital, Copenhagen, Denmark, [4]Department of Oncology, Wayne State University, Detroit, USA, [5]Department of Clinical Biochemistry, Gentofte Hospital, [6]Department of Clinical Biochemistry, Herlev Hospital, Copenhagen, Denmark, [7]Department of Human Genetics, University of Michigan, Ann Arbor, USA and [1–3,5–6]Copenhagen University Hospital and Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark.

*Corresponding author. Department of Clinical Biochemistry, Rigshospitalet, Copenhagen University Hospital, Blegdamsvej 9, DK-2100, Copenhagen, Denmark. E-mail: ruth.frikke-schmidt@regionh.dk

Accepted 9 October 2014

Abstract

Background The aetiology of ischaemic heart disease (IHD) is complex and is influenced by a spectrum of environmental factors and susceptibility genes. Traditional statistical modelling considers such factors to act independently in an additive manner. The Patient Rule-Induction Method (PRIM) is a multi-model building strategy for evaluating risk attributable to context-dependent gene and environmental effects.

Methods PRIM was applied to 9073 participants from the prospective Copenhagen City Heart Study (CCHS). Gender-specific cumulative incidences were estimated for subgroups defined by categories of age, smoking, hypertension, diabetes, body mass index, total cholesterol, high-density lipoprotein cholesterol and triglycerides and by 94 single nucleotide variants (SNVs).Cumulative incidences for subgroups were validated using an independently ascertained sample of 58 240 participants from the Copenhagen General Population Study (CGPS).

Results In the CCHS the overall cumulative incidences were 0.17 in women and 0.21 in men. PRIM identified six and four mutually exclusive subgroups in women and men, respectively, with cumulative incidences of IHD ranging from 0.02 to 0.34. Cumulative incidences of IHD generated by PRIM in the CCHS were validated in four of the six subgroups of women and two of the four subgroups of men in the CGPS.

Conclusions PRIM identified high-risk subgroups characterized by specific contexts of selected values of traditional risk factors and genetic variants. These subgroups were validated in an independently ascertained cohort study. Thus, a multi-model strategy may identify groups of individuals with substantially higher risk of IHD than the overall risk for the general population.

Key words: Cardiovascular disease, risk factor, genetic epidemiology

Key Messages

- The present application of the Patient Rule-Induction Method (PRIM) for IHD shows that PRIM is able to identify high-risk subgroups of individuals characterized by selected values of traditional risk factors and candidate genetic variants.
- These findings are novel, and suggest that a multi-model strategy is able to identify groups of individuals characterized by specific contexts with substantially higher risk of IHD than the overall risk for the general population.
- This result has clinical relevance because such high-risk subgroups may benefit the most from aggressive preventive treatments.

Introduction

Ischaemic heart disease (IHD) is the leading cause of morbidity and mortality worldwide.[1] The aetiology of IHD is complex and is influenced by a spectrum of environmental factors and susceptibility genes.[2] Traditional statistical modelling considers such factors to act independently in an additive manner, and assumes that the expected relationship between disease status and variation in genetic and environmental risk factors is the same for all individuals in the population under study. This perspective does not take into account the fact that the effects of a particular genetic variant on an individual's risk of disease may depend on context, defined by established environmental risk factors, and by the background genotype.[3,4]

Currently, over 1800 genome-wide association studies (GWAS) have reported validated associations between common single nucleotide variants (SNVs) and complex disorders including cardiovascular disease, cancer, diabetes and psychiatric diseases [(http//:www.genome.gov/gwas-studies]. As these diseases are common, they place the greatest public health burden on society.[4] However, in every case substantial heritable variation in risk of disease and in biological risk factors for disease is not explained by common SNVs identified by the GWAS.[2,5–8] Possible explanations for this 'missing heritability' include rare variant effects, the effects of gene-gene and gene-environment interactions, and aetiological heterogeneity (i.e. different combinations of genes and environments influence risk in different subgroups of the population) that are not considered by GWAS.[5,9] The systematic mapping of regions of transcription, transcription factor association, chromatin structure and histone modification, recently published by the Encyclopedia of DNA Elements (ENCODE) project, revealed that the function of the human transcriptional regulatory network is highly context-specific.[10,11]

Analytical strategies to investigate the context-dependent effects of genomic variations on the risk of a common disease having a complex multifactorial aetiology are currently in their infancy.[2,5,9] The Patient Rule-Induction method (PRIM)[12–15] is a model-building strategy for evaluating risk that acknowledges context-dependent gene and environmental effects and aetiological heterogeneity whereby different combinations of genetic and environmental risk factors are predictive of disease outcome in different genetic and environmentally defined subgroups of the population. This strategy makes possible the identification of combinations of risk factor values, environmental strata and/or genetic variants that characterize mutually exclusive subgroups of individuals that differ in average risk as measured by the cumulative incidence of the disease of interest.

In this paper we consider traditional IHD risk factors and 94 SNVs in 22 candidate genes in an application of PRIM to modelling the cumulative incidence of IHD in subgroups of a population-based sample of 9073 individuals enrolled in the prospective Copenhagen City Heart Study (CCHS). We validated the resultant models in an independently ascertained cohort of 58 240 individuals from the Copenhagen General Population Study (CGPS).

Methods

Participants

Studies were approved by institutional review boards and Danish ethical committees (KF-100.2039/91; KF-01-144/01; H-KF-01-144/01), and conducted according to the Declaration of Helsinki. Informed consent was obtained from participants. All participants were White and of Danish descent, as determined by the Danish Central Person Registration System. Each individual was included in only one of the two studies.

The Copenhagen City Heart Study (CCHS)

This is a prospective study of the general population initiated in 1976–78 with follow-up examinations in 1981–83, 1991–94 and 2001–03.[16–18] Individuals were randomly selected based on the Danish Central Person Registration System to reflect the adult Danish general population aged 20–100 years. Data were obtained from a questionnaire, a physical examination and blood samples. Participants in the present study were from the 1991–94 examination, when blood samples for DNA extraction were drawn.[19] Of the 16 563 individuals invited, 10 135 participated (61% response rate). Among these, we included 9073 individuals in the present study (5151 women and 3922 men) with no prior history of IHD and with complete clinical and laboratory data available. Follow-up started at the 1991–94 examination and ended at the occurrence of an IHD event, date of death, emigration or on 10 May 2010 (last update of registry), whichever came first. Follow-up was up to 19 years, and was 100% complete, i.e. none was lost to follow-up. We used this sample to apply the PRIM method to model the context dependency of genetic and environmental effects.

The Copenhagen General Population Study (CGPS)

This is a prospective study of the Danish general population initiated in 2003 with ongoing enrolment.[16–18] Data were obtained from a questionnaire, a physical examination and blood samples including DNA extraction. We included 58 240 participants with complete clinical and laboratory data available and with no prior history of IHD. Follow-up began at study entry and ended at the occurrence of an IHD event, death, emigration or on 10 May 2010 (last update of registry), whichever came first. Follow-up was up to 7 years, and was 100% complete, i.e. none was lost to follow-up. We used this sample to validate the context-dependent models estimated using the CCHS sample.

Definition of disease endpoint

Information on diagnoses of IHD (WHO International Classification of Diseases, ICD8:410–414; ICD10:I20–I25) was collected from 1977 through 2010 by reviewing all hospital admissions and diagnoses entered in the national Danish Patient Registry and all causes of death entered in the national Danish Causes of Death Registry as described.[16–18] Ischaemic heart disease was defined by fatal or non-fatal myocardial infarction or characteristic symptoms of angina pectoris, including revascularization procedures,[20] as detailed in international guidelines.[20–22]

Laboratory analyses

Total cholesterol, high-density lipoprotein (HDL) cholesterol and triglycerides were measured by colourimetric assays (Boehringer Mannheim GmbH, Mannheim, Germany, and Konelab, Helsinki, Finland).

Genotyping

Genotyping was by TaqMan assays using the ABI PRISM 7900HT Sequence Detection System (Applied Biosystems, Foster City, CA, USA) or by restriction enzyme assays. Genotypes were verified by genotyping at least 50 randomly selected samples of each variant by two different methods (TaqMan assays plus sequencing or restriction enzyme assay); there was 100% agreement between TaqMan and sequencing/restriction enzyme assay results. Call rates for genotypes were above 99.9% for all assays, due to reruns.

Definitions of traditional risk factors

Age was stratified in three groups, <45, 45–65 and ≥65 years. The definitions of smoking, hypertension and diabetes have been described previously.[12] In brief, each risk factor was dichotomized to define a high-risk group; a history of smoking (current smoker at any examination), a history of hypertension (systolic blood pressure ≥140 mmHg, diastolic blood pressure ≥90 mmHg, use of antihypertensive drugs, or any combination of these at any examination) or a history of diabetes (self-reported disease, use of insulin, use of oral hypoglycaemic drugs, non-fasting plasma glucose ≥11.1 mmol/l at any examination). Body mass index was categorized according to World Health Organization definitions. Recommendations of the National Cholesterol Education Program Expert Panel, National Institutes of Health, were used to define dyslipidaemic sub-groups (National Cholesterol Education Program, National Heart, Lung, and Blood Institute 2002). Dyslipidaemia was diagnosed when total cholesterol was ≥5.18 mmol/l (200 mg/dl), HDL cholesterol was <1.04 mmol/l (40 mg/dl) or triglycerides was ≥1.70 mmol/l (150 mg/dl).

Statistical analyses

Patient Rule-Induction Method

The PRIM was introduced by Friedman and Fisher[23] and modified by Dyson *et al.*[12,13] and Dyson and Sing[14] for use in identifying mutually exclusive subgroups of individuals with varying cumulative incidences of IHD. The subgroups are defined by terms (selected values of predictor variables) and are created through repeated implementations of the peeling and pasting algorithms. Peeling is an iterative process that creates a subgroup by excluding individuals with particular values of predictor variables, whereas pasting iteratively amends individuals to the subgroup, also based upon values of predictor variables, after the peeling stage has been completed. We applied PRIM to obtain risk factor models for mutually exclusive subgroups of risk separately for women and men from the CCHS. We then applied these models of risk to the prediction of IHD in the CGPS sample to evaluate the validation of the models.

Validation analyses

A direct comparison of the value of the cumulative risk estimated for each subgroup modelled by PRIM using the CCHS sample with the observed cumulative incidence for the corresponding subgroup in the CGPS is inappropriate because the samples have different overall cumulative incidences (21.0% in men and 16.7% in women for CCHS and 3.2% in men and 1.9% in women for CGPS). Therefore, we rescaled the estimate of the cumulative incidence of each subgroup predicted by the PRIM models as a deviation from the cumulative incidence in the total sample divided by the cumulative incidence in the total sample. The computed confidence intervals were likewise rescaled. Rescaling was necessary to adjust for differences in IHD incidence in the training and validation data sets to allow for meaningful evaluation of the prediction models created. We used this percent change metric rather than an odds ratio or relative risk to better manage the boundary effects that exist with odds ratios or relative risk computations. We used the same rescaling transformation for the observed cumulative incidence for each subgroup in the CGPS defined by PRIM analysis of the CCHS sample. If the rescaled observed cumulative incidence in the CGPS subgroup, defined by the PRIM modelling of the CCHS sample, falls within the 95% rescaled predicted confidence interval for the estimate of the cumulative incidence for the subgroup in the CCHS sample, we concluded that the subgroup was validated.

Results

Characteristics

Characteristics of the 5151 women and 3922 men from the Copenhagen City Heart Study are shown in Table 1. The distribution of individuals in categories of age, smoking, hypertension, diabetes, body mass index, total cholesterol, HDL cholesterol and triglycerides as well as cumulative incidences of IHD differed as a function of gender (Table 1). Men were more frequently smokers, hypertensives or diabetics, and more frequently had body mass index ≥25 kg/m^2, HDL cholesterol <1.04 mmol/l (40 mg/dl) and triglycerides ≥1.70 mmol/l (150 mg/dl) compared with women. Women more frequently had total cholesterol ≥5.18 mmol/l (200 mg/dl) compared with men (χ^2: P <0.0001). These differences in the cumulative incidence of IHD and the frequencies of risk factors between genders justify carrying out model-building separately in women and men.

Table 1. Characteristics of individuals from the Copenhagen City Heart Study

Characteristics	Women (N=5151)	Men (N=3922)	*P*-value*	95% CI**
Cumulative incidence of IHD (%)	861 (16.7%)	824 (21.0%)	1.9×10^{-7}	−5.9 to −2.7
Age ≥65 years (%)	2107 (40.9%)	1282 (32.7%)	1.1×10^{-15}	6.2 to 10.2
Smoking (%)	2883 (56.0%)	2582 (65.8%)	1.9×10^{-21}	−11.9 to −7.9
Hypertension (%)	2919 (56.7%)	2586 (65.9%)	3.5×10^{-19}	−11.3 to −7.3
Diabetes (%)	156 (3.0%)	230 (5.9%)	3.4×10^{-11}	−3.7 to −2.0
Body mass index ≥25 kg/m^2 (%)	2284 (44.3%)	2216 (56.5%)	1.3×10^{-36}	−14.2 to −10.1
Total cholesterol ≥5.18 mmol/l (%)	4114 (79.9%)	2906 (74.1%)	7.5×10^{-11}	4.0 to 7.5
HDL cholesterol <1.04 mmol/l (%)	280 (5.4%)	798 (20.4%)	7.7×10^{-105}	−16.3 to −13.5
Triglycerides ≥1.70 mmol/l (%)	1875 (36.4%)	1921 (49.0%)	2.4×10^{-33}	−14.6 to −10.5

Smoking, current-smoker at any examination; hypertension, systolic blood pressure ≥140 mmHg, diastolic blood pressure ≥90 mmHg, use of antihypertensive drugs, or any combination of these at any examination; diabetes, self-reported disease, use of insulin, use of oral hypoglycaemic drugs, non-fasting plasma glucose ≥11.1 mmol/l, or any combination of these at any examination. Body mass index was categorized according to World Health Organization definitions. Recommendations of the National Cholesterol Education Program Expert Panel, National Institutes of Health, were used to define dyslipidaemic subgroups (National Cholesterol Education Program, National Heart, Lung, and Blood Institute 2002). To convert total cholesterol and HDL cholesterol to mg/dl, divide by 0.0259; to convert triglycerides to mg/dl, divide by 0.0113. HDL, high-density lipoprotein; IHD, ischaemic heart disease.

*P-value using χ^2 tests.

**The confidence interval (CI) for the difference in risk factor frequencies between women and men is calculated using the normal approximation to the binomial distribution.

Candidate genetic variants

Table 2 describes the94 SNVs considered in the PRIM analysis of the CCHS sample. These SNVs have previously been described in the literature as candidate variants for IHD and for IHD risk factors. Minor allele frequencies ranged from 0.01 to 49%. Frequencies of the genotypes of 92 out of 94 genetic variants did not deviate from the Hardy-Weinberg expectations (P-values $\geq 5 \times 10^{-2}$). Frequencies of the genotypes of each of the two remaining variants (*ABCA1* E1172D and *APOA1*-150G>A) deviated significantly from the Hardy–Weinberg expectations (P-value $= 10^{-2}$ and 10^{-2}, respectively). Sequencing of the heterozygotes and homozygotes for the minor allele of *ABCA1* E1172D and *APOA1*-150G>A confirmed the original genotype calls. Hence, no technical reasons for the deviations from Hardy–Weinberg expectations were detected.

PRIM-defined subgroups from the CCHS sample

The risk factor categories (see Methods) were simultaneously considered with the 94 SNVs (Table 2) in building subgroups using the PRIM in each gender separately. Five and three mutually exclusive subgroups were defined in women and men, respectively, as well as a final remainder subgroup for each gender.

In the sample of women, the cumulative incidences of IHD in the five sequentially identified subgroups and one final remainder subgroup (summarized in Figure 1) were: 0.31 [95% confidence interval (CI): 0.28-0.34] for age ≥65 years and *LIPC* N193S AG+GG and hypertension; 0.23 (95% CI: 0.21–0.25) for age ≥65 or diabetes; 0.15 (95% CI: 0.13–0.18) for hypertension and age <65 years and no diabetes; 0.12 (95% CI: 0.09–0.15) for triglycerides ≥1.70 mmol/l (150 mg/dl) and age <65 years and no diabetes and no hypertension; 0.12 (95% CI: 0.07–0.15) for age 45–64 years and smoking and triglycerides <1.70 mmol/l (150 mg/dl) and no diabetes and no hypertension; and 0.02 (95% CI: 0.02–0.04) for the remainder subgroup.

In the sample of men, the cumulative incidences of IHD in the three sequentially identified subgroups and one final remainder subgroup (summarized in Figure 2) were: 0.34 (95% CI: 0.31–0.36) for age ≥65 years and hypertension or diabetes; 0.26 (95% CI: 0.24–0.30) for body mass index ≥25 kg/m^2 and triglycerides ≥1.70 mmol/l (150 mg/dl) and total cholesterol ≥5.18 mmol/l (200 mg/dl) and no diabetes; 0.16 (95% CI: 0.14–0.18) for age ≥45 years and no diabetes; and 0.04 (95% CI: 0.03–0.08) for the remainder subgroup.

Validation of CCHS subgroups in CGPS

Figures 3 and 4 present the rescaled cumulative incidences and their 95% CIs, for the PRIM subgroups of the CCHS samples of women and men, respectively, shown as blue diamonds with confidence intervals. The rescaled cumulative incidences for each of the subgroups in the CGPS sample, defined by the PRIM modelling of the CCHS sample, are denoted by arrows. The PRIM model for IHD risk was validated in four of the six subgroups in women, and in two of the four subgroups in men. Risk was smallest in those subgroups that did not validate. The highest risk in both genders in both samples was associated with subgroups including individuals aged ≥65 years and with hypertension and/or diabetes.

The distribution of risk factor categories among subgroups in CCHS and CGPS is presented in Table 3 for women and in Table 4 for men. The percentage of individuals in the high-risk category was largely comparable between CCHS and CGPS for each subgroup, except for total cholesterol, where individuals in the CCHS more frequently than individuals in the CGPS had total cholesterol ≥5.18 mmol/l (200 mg/dl).

Discussion

The main finding of the present study is that the PRIM identifies high-risk subgroups of individuals characterized by selected values of traditional risk factors and candidate genetic variants. The majority of these subgroups had cumulative incidences of similar relative magnitude in an independently ascertained cohort study of the same population of inference; these findings are novel, and suggest that a multi-model strategy is able to identify groups of individuals characterized by specific contexts with substantially higher risk of IHD than the overall risk for the general population. This result has clinical relevance because such high-risk subgroups may benefit the most from aggressive preventive treatments.

Extensive studies of model organisms have established that the phenotypic effects of single loci are contingent on the contexts defined by other loci (gene-gene interactions) and the history of exposures to environmental influences (gene-environmental interactions).[24–26] It has been argued that context-dependent genetic effects are the rule rather than the exception in explaining the biological aetiology of a human trait that has a complex multifactorial aetiology.[3] However, the estimation and testing of gene-gene interactions and gene-environmental interactions using traditional statistical methods to model phenotypic variation of such traits has proved to be inappropriate when using non-experimental data collected in population-based studies.[27] This is of special concern in studies of the common human diseases that have a complex multifactorial aetiology.[3,28–30] The primary reasons include that: (i) correlations between loci are inherent in non-experimental

Table 2. Candidate single nucleotide variants (N=94) genotyped in the Copenhagen City Heart Study (N=9073)

Gene	Rs number	Amino acid substitution[a]	Alternative literature name	Minor allele (frequency %)	Hardy-Weinberg equilibrium *P*-value
ABCA1	Rs2740483	–	−99G>C	C (31.2)	0.20
	Rs1800977	–	−14C>T	T (34.5)	0.25
	Rs111292742	–	+35C>G	G (4.1)	0.71
	Rs2230806	R219K	–	A (26.5)	0.39
	–	S364C	–	G (0.01)	1.00
	Rs9282543	V399A	–	C (0.2)	0.85
	Rs2066718	V771M	–	A (3.3)	0.84
	Rs35819696	T774P	–	C (0.2)	0.88
	Rs138880920	K776N	–	C (0.2)	0.87
	Rs2066715	V825M	–	A (5.7)	0.17
	Rs2066714	I883M	–	G (12.1)	0.19
	–	P1065S	–	T (0.01)	1.00
	Rs33918808	E1172D	–	C (2.8)	0.01
	–	G1216V	–	T (0.01)	0.99
	Rs2230808	R1587K	–	A (23.8)	0.20
	Rs146292819	N1800H	–	C (0.1)	0.91
	–	R2144X	–	T (0.01)	0.99
ACE	Rs1799752	–	I/D or D/I	I (49.0)	0.60
AGT	Rs5050	–	–	C (16.1)	0.58
	Rs5051	–	–	A (40.0)	0.55
	Rs4762	T207M	T174M	T (12.2)	0.85
	Rs699	M268T	M235T	C (40.4)	0.35
ANGPTL4	Rs116843064	E40K	–	A (3.0)	0.53
APOA1	–	–	−647A>G	G (0.5)	0.13
	Rs12718466	–	−560A>C	C (3.5)	0.73
	Rs670	–	−310G>A	A (16.1)	0.68
	Rs5069	–	−151C>T	T (3.5)	0.70
	Rs1799837	–	−150G>A	A (0.7)	0.01
	–	–	*141G>A	A (0.2)	0.89
APOB	Rs1367117	T98I	T71I	T (33.4)	0.25
	Rs531819	–	–	A (13.7)	0.75
	Rs679899	A618V	A591V	T (46.8)	0.47
	Rs10199768	–	–	C (29.3)	0.69
	Rs3791980	–	–	A (44.9)	0.97
	Rs693	T2515T	T2488T	T (48.0)	0.33
	Rs676210	P2739L	P2712L	T (20.5)	0.70
	–	R3507P	R3480P	C (0.02)	0.99
	Rs5742904	R3527Q	R3500Q	A (0.03)	0.98
	Rs12713559	R3558C	R3531C	T (0.04)	0.97
	Rs1801701	R3638Q	R3611Q	A (9.5)	0.19
	–	R4046W	R4019W	T (0.02)	0.99
	Rs1042031	E4181K	E4154K	A (17.5)	0.21
	Rs1042034	–	–	G (20.5)	0.74
APOE	Rs449647	–	−491A>T	T (15.4)	0.69
	Rs769446	–	−427T>C	C (10.8)	0.29
	Rs405509	–	−219G>T	T (47.3)	0.80
	Rs429358	C130R	Epsilon 4/C112R	C (16.6)	0.86
	Rs7412	R176C	Epsilon 2/R158C	T (8.2)	0.06
APOJ/CLU	Rs714787	–	–	A (9.6)	0.39
	Rs7012217	–	–	G (29.9)	0.92
	Rs9331936	N317H	N369H	C (0.01)	0.99
	Rs9331949	–	–	G (2.6)	0.86

(Continued)

Table 2. Continued

Gene	Rs number	Amino acid substitution[a]	Alternative literature name	Minor allele (frequency %)	Hardy-Weinberg equilibrium *P*-value
Chr9	Rs10757274	–	–	G (44.9)	0.11
	Rs2383206	–	–	G (47.3)	0.05
F5	Rs6025	R506Q	R534Q/Factor V Leiden	G (3.9)	0.85
F2	Rs1799963	–	G20210A	A (1.1)	0.98
FGB	Rs1800790	–	−455G>A	A (20.0)	0.60
HFE	Rs1799945	H63D	–	G (12.7)	0.40
	Rs1800562	C282Y	–	A (5.6)	0.24
LIPC	Rs1800588	–	−480C>T	T (21.3)	0.90
	Rs6078	V95M	V73M	A (3.0)	0.83
	Rs6083	N215S	N193S	G (37.2)	0.30
	Rs121912502	S289F	S267F	T (0.4)	0.73
	Rs3829462	L356F	L334F	C (1.7)	0.11
	Rs28933094	T405M	T383M	T (0.3)	0.82
ITGB3	Rs5918	L59P	L33P	C (16.4)	0.92
	Rs36080296	L66R	L40R	G (0.3)	0.76
LDLR	–	W44X	W23X	A (0.01)	1.00
	–	W87G	W66G	G (0.02)	0.99
	Rs11669576	A391T	A370T	A (5.2)	0.34
	–	W577S	W556S	C (0.01)	1.00
LPL	Rs1800590	–	−93T>G	G (1.5)	0.94
	–	–	−53G>C	C (0.1)	0.92
	Rs1801177	D36N	D9N	A (1.4)	0.59
	Rs118204057	G215E	G188E	A (0.03)	0.98
	Rs268	N318S	N291S	G (2.5)	0.79
	Rs328	S474X	S447X	G (9.9)	0.11
MTHFR	Rs1801133	A222V	C677T	T (30.8)	0.43
PCSK9	Rs11591147	R46L	–	T (1.2)	0.82
SERPINA1	Rs17580	E288V	E264V/S-allele	T (2.9)	0.12
	Rs28929474	E366K	E342K/Z-allele	A (2.8)	0.86
SOD3	Rs1799895	R231G	R213G	C (1.1)	0.93
ZNF202	Rs10736530	–	−685G>A	A (3.8)	0.32
	Rs10893081	–	−660A>G	G (30.2)	0.82
	–	–	−447T>C	C (0.4)	0.70
	–	–	−232C>T	T (0.1)	0.91
	–	–	−122C>T	T (0.2)	0.83
	–	–	−118G>T	T (5.2)	0.31
	Rs2272142	–	+34G>A	A (1.4)	0.32
	Rs1144507	A154V	–	T (30.0)	0.54
	Rs61767139	K259E	–	G (1.0)	0.86
	–	V274L	–	C (0.01)	1.00
	–	R605W	–	T (0.04)	0.97
	Rs3183878	–	*2T>G	G (39.0)	0.48

ABCA1, ATP-binding-cassette transporter A1; ACE, angiotensinogen converting enzyme; AGT, angiotensinogen; ANGPTL4, angiopoetin-like4; APOA1, apolipoprotein A1; APOB, apolipoprotein B; APOE, apolipoprotein E; APOJ, apolipoprotein J (clusterin); Chr9, chromosome 9; CLU, clusterin; F5, factor V; F2, factor II (prothrombin); FGB, fibrinogen β polypeptide; HFE, haemochromatosis gene; LIPC, hepatic lipase; ITGB3, integrin beta 3 (platelet glycoprotien IIb/IIIa); LDLR, low-density lipoprotein receptor; LPL, lipoprotein lipase; MTHFR, methylenetetrahydrofolate reductase; PCSK9, proprotein convertase, subtilisin/kexin-type 9; SERPINA1, alfa 1-antitrypsin; SOD3, superoxide dismutase 3, extracellular; ZNF202, zinc finger protein 202.

[a]Nomenclature according to den Dunnen *et al.*[31]

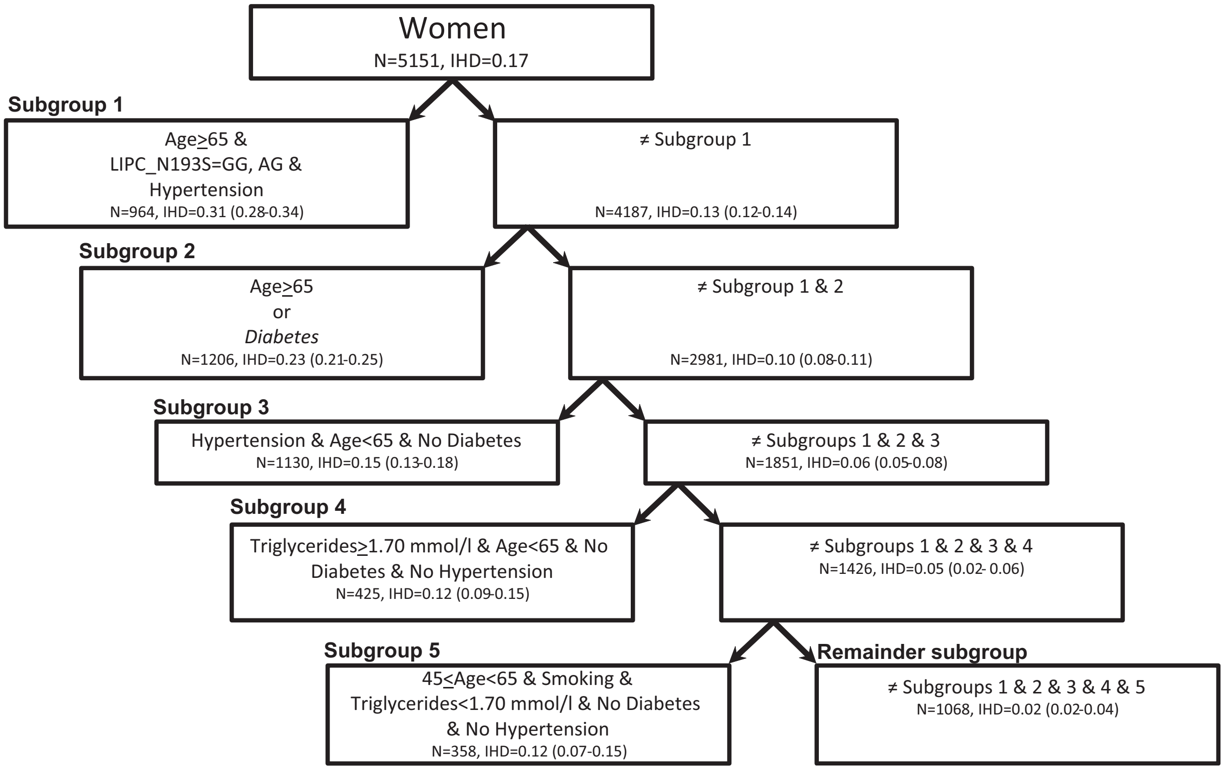
Women
N=5151, IHD=0.17
Subgroup 1
Age≥65 &
LIPC_N193S=GG, AG &
Hypertension
N=964, IHD=0.31 (0.28-0.34)
≠ Subgroup 1
N=4187, IHD=0.13 (0.12-0.14)
Subgroup 2
Age≥65
or
Diabetes
N=1206, IHD=0.23 (0.21-0.25)
≠ Subgroup 1 & 2
N=2981, IHD=0.10 (0.08-0.11)
Subgroup 3
Hypertension & Age<65 & No Diabetes
N=1130, IHD=0.15 (0.13-0.18)
≠ Subgroups 1 & 2 & 3
N=1851, IHD=0.06 (0.05-0.08)
Subgroup 4
Triglycerides≥1.70 mmol/l & Age<65 & No Diabetes & No Hypertension
N=425, IHD=0.12 (0.09-0.15)
≠ Subgroups 1 & 2 & 3 & 4
N=1426, IHD=0.05 (0.02- 0.06)
Subgroup 5
45≤Age<65 & Smoking &
Triglycerides<1.70 mmol/l & No Diabetes
& No Hypertension
N=358, IHD=0.12 (0.07-0.15)
Remainder subgroup
≠ Subgroups 1 & 2 & 3 & 4 & 5
N=1068, IHD=0.02 (0.02-0.04)

Figure 1. Consecutively identified, mutually exclusive subgroups with decreasing cumulative incidences of ischaemic heart disease using PRIM in women from the Copenhagen City Heart Study. The data set contained 5151 women, with an overall cumulative incidence of ischaemic heart disease of 0.17. The first subgroup was defined by three peeling terms (Age≥65 & LIPC_N193S = GG, AG & Hypertension). The process of producing a new subgroup based on the unassigned individuals from the previous partition continued until all individuals were assigned to a subgroup. The individuals that were not included in any of the subgroups were assigned to the remainder subgroup. IHD, cumulative incidence of ischaemic heart disease; *LIPC*_N193S = rs6083 variant in the hepatic lipase gene (Table 2). Parentheses after IHD indicate 95% confidence interval for the cumulative incidence. Age is in years. Terms in italics are the result of a paste operation. ≠, different from.

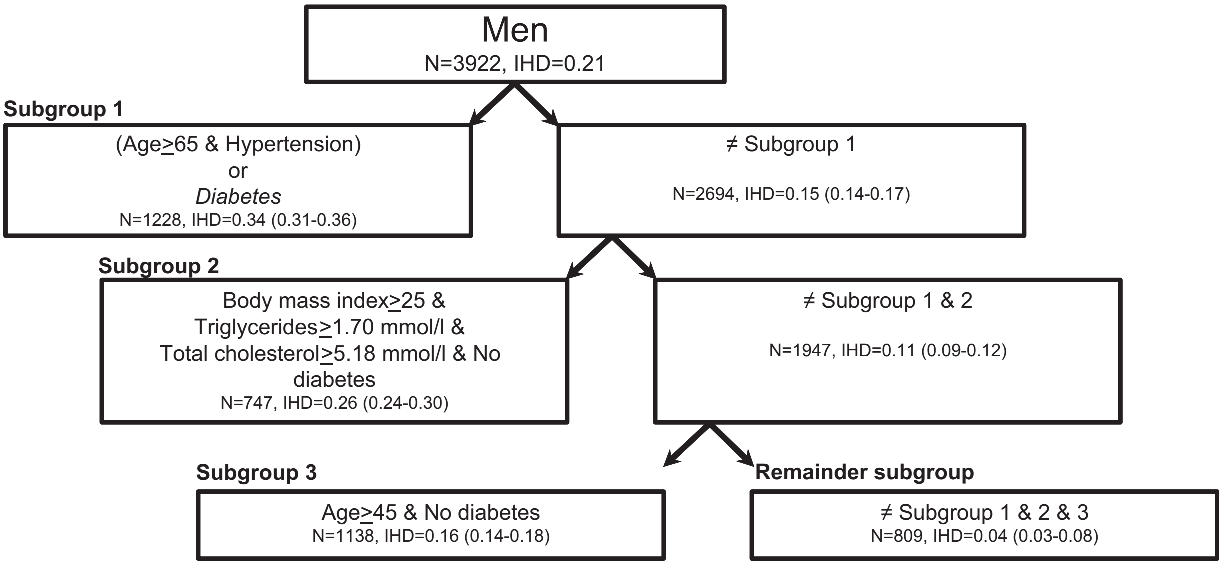
Men
N=3922, IHD=0.21
Subgroup 1
(Age≥65 & Hypertension)
or
Diabetes
N=1228, IHD=0.34 (0.31-0.36)
≠ Subgroup 1
N=2694, IHD=0.15 (0.14-0.17)
Subgroup 2
Body mass index≥25 &
Triglycerides≥1.70 mmol/l &
Total cholesterol≥5.18 mmol/l & No diabetes
N=747, IHD=0.26 (0.24-0.30)
≠ Subgroup 1 & 2
N=1947, IHD=0.11 (0.09-0.12)
Subgroup 3
Age≥45 & No diabetes
N=1138, IHD=0.16 (0.14-0.18)
Remainder subgroup
≠ Subgroup 1 & 2 & 3
N=809, IHD=0.04 (0.03-0.08)

Figure 2. Consecutively identified, mutually exclusive subgroups with decreasing cumulative incidences of ischaemic heart disease using PRIM in men from the Copenhagen City Heart Study. The data set contained 3922 men, with an overall cumulative incidence of ischaemic heart disease of 0.21. The first subgroup was defined by two peeling terms (Age≥65 & Hypertension) and one pasting term (Diabetes). The process of producing a new subgroup based on the unassigned individuals from the previous partition continued until all individuals were assigned to a subgroup. The individuals that were not included in any of the subgroups were assigned to the remainder subgroup. IHD, cumulative incidence of ischaemic heart disease. Parentheses after IHD indicate 95% confidence interval for the cumulative incidence. Age is in years. Terms in italics are the result of a paste operation. ≠, different from.

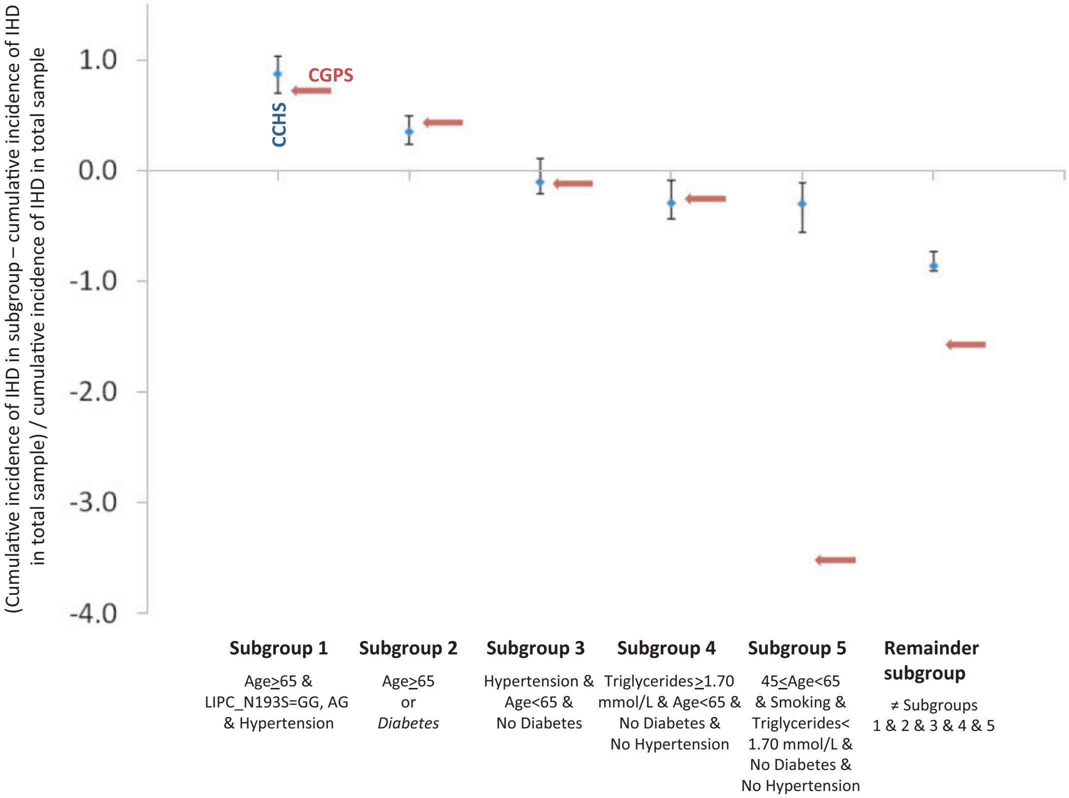

Figure 3. Validation of CCHS PRIM model for women in the CGPS sample. To compare cumulative incidences between subgroups from two different population samples with different overall cumulative incidences, for each subgroup we used its deviation from the cumulative incidence in the total sample divided by the cumulative incidence in the total sample. Estimates for CCHS are indicated as blue diamonds with 95% confidence intervals, and CGPS estimates appear as red arrows. CCHS, Copenhagen City Heart Study; CGPS, Copenhagen General Population Study; IHD, ischaemic heart disease. Age is in years. Terms in italics are the result of a paste operation.

observational studies and confound the phenotypic effect of a single locus with phenotypic effects of correlated loci; and (ii) aetiological heterogeneity, i.e. different subsets of interacting genetic and environmental risk factors are responsible for determining the complex trait phenotype in different subgroups of the population at risk. The application of PRIM in our study addresses both of these issues by focusing on estimating the number of subgroups, each with a particular combination of risk factors, that best explains the distribution of risk of IHD in a population-based sample. In this way the resultant PRIM models incorporate the combined influences of context-dependent gene and environmental effects and aetiological heterogeneity. And most important, rather than estimating interactions with standard statistical procedures and focusing on the causes of heritability in the sampled population, this analytical strategy produces combinations of interacting risk factors which may have utility in informing medicine in achieving improved prevention and treatment selection.

An even more difficult issue in the characterization of the role of interacting risk factors is the widely accepted practice of validation of a hypothesized prediction model, developed in one study, in subsequent studies of independently ascertained samples from different geographical and ethnically diverse populations.[28] An optimal replicate sample for a test of model validation should be similar to the model-building sample with respect to ethnicity, genetic structure, relative frequencies of risk factors and access to medical care. Our study is unique in addressing these criteria by considering validation in a sample that is representative of the population being studied: the only exception being that the sample ascertained from the CGPS included adjacent geographical regions north and west of the city of Copenhagen from which the CCHS sample was drawn. Ethnic homogeneity was assured through access to information in the national Central Person Register that facilitated the inclusion of individuals whose Danish ancestry was documented by the place of birth of ancestors three generations back. Similar access to medical care was assured by access of all participants to the tax-financed and centrally administered primary, secondary and tertiary care offered by the Danish Health System. Finally, risk factor frequencies were largely comparable for each subgroup between the two studies, with the exception of levels of

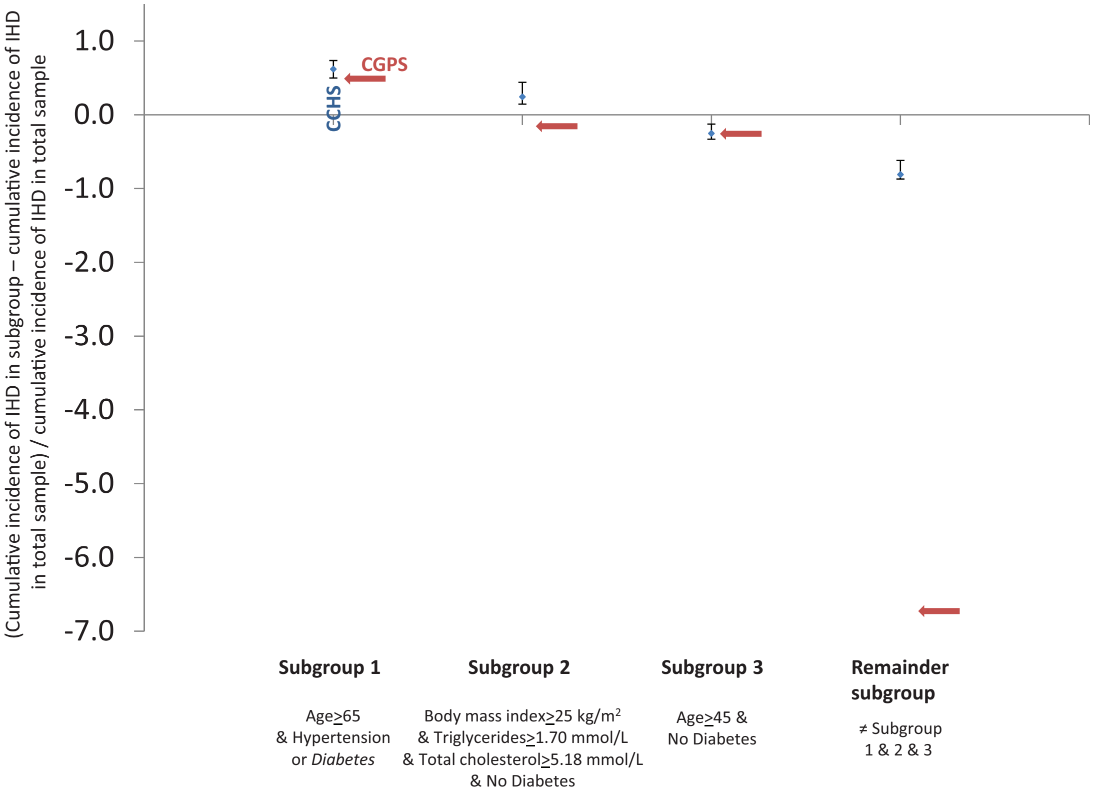

Figure 4. Validation of CCHS PRIM model for men in the CGPS sample. To compare cumulative incidences between subgroups from two different population samples with different overall cumulative incidences, for each subgroup we used its deviation from the cumulative incidence in the total sample divided by the cumulative incidence in the total sample. Estimates for CCHS are indicated as blue diamonds with 95% confidence intervals, and CGPS estimates appear as red arrows. CCHS, Copenhagen City Heart Study; CGPS, Copenhagen General Population Study; IHD, ischaemic heart disease. Age is in years. Terms in italics are the result of a paste operation.

Table 3. Comparisons of the distribution of traditional risk factors among subgroups in women in the Copenhagen City Heart Study and the Copenhagen General Population Study

Risk factors	Subgroup 1		Subgroup 2		Subgroup 3		Subgroup 4		Subgroup 5		Remainder subgroup	
	CCHS N=964	CGPS N=3723	CCHS N=1206	CGPS N=5851	CCHS N=1130	CGPS N=9586	CCHS N=425	CGPS N=2629	CCHS N=358	CGPS N=2219	CCHS N=1068	CGPS N=8866
Cumulative incidence of IHD (%)	31.3	3.9	22.6	3.3	15.0	1.7	11.8	1.5	11.8	0.4	2.3	0.7
Age ≥65 years (%)	100.0	100.0	94.8	99.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Smoking (%)	51.2	54.1	54.5	56.8	60.7	57.0	68.0	66.8	100.0	100.0	37.4	43.1
Hypertension (%)	100.0	100.0	68.4	65.4	100.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0
Diabetes (%)	4.9	6.3	9.0	11.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Body mass index ≥25 kg/m^2 (%)	55.6	60.5	49.9	53.9	54.7	56.1	53.2	55.4	28.8	26.9	18.6	29.5
Total cholesterol ≥5.18 mmol/l (%)	93.9	80.0	91.2	76.8	86.6	74.4	80.9	79.2	81.6	34.1	46.4	55.4
HDL cholesterol <1.04 mmol/l (%)	5.4	4.0	6.6	5.2	5.3	5.1	12.0	12.7	3.1	3.3	2.4	1.8
Triglycerides ≥1.70 mmol/l (%)	45.0	38.7	45.3	35.7	41.6	31.9	100.0	100.0	0.0	0.0	0.0	0.0

Smoking, current-smoker at any examination; hypertension, systolic blood pressure ≥140 mmHg, diastolic blood pressure ≥90 mmHg, use of antihypertensive drugs, or any combination of these at any examination; diabetes, self-reported disease, use of insulin, use of oral hypoglycaemic drugs, non-fasting plasma glucose ≥11.1 mmol/l or any combination of these at any examination. Body mass index was categorized according to World Health Organization definitions. Recommendations of the National Cholesterol Education Program Expert Panel, National Institutes of Health, were used to define dyslipidaemic subgroups (National Cholesterol Education Program, National Heart, Lung, and Blood Institute 2002). To convert total cholesterol and HDL cholesterol to mg/dl, divide by 0.0259; to convert triglycerides to mg/dl, divide by 0.0113. CCHS, Copenhagen City Heart Study; CGPS, Copenhagen General Population Study; HDL, high-density lipoprotein; IHD, ischaemic heart disease.

Table 4. Comparisons of the distributions of traditional risk factors among subgroups in men in the Copenhagen City Heart Study and the Copenhagen General Population Study

Risk factors	Subgroup 1		Subgroup 2		Subgroup 3		Remainder subgroup	
	CCHS N=1228	CGPS N=6390	CCHS N=747	CGPS N=5664	CCHS N=1138	CGPS N=9643	CCHS N=809	CGPS N=3669
Cumulative incidence of IHD (%)	34.0	6.1	26.1	2.7	15.7	2.5	4.0	0.4
Age ≥65 years (%)	90.8	91.6	5.2	5.5	11.2	10.9	0.0	0.0
Smoking (%)	65.6	74.1	67.7	65.1	73.7	62.0	53.3	43.5
Hypertension (%)	97.8	97.8	71.1	64.3	54.4	50.4	29.1	40.1
Diabetes (%)	18.7	17.4	0.0	0.0	0.0	0.0	0.0	0.0
Body mass index≥25 kg/m^2 (%)	65.9	72.5	100.0	100.0	41.2	50.8	23.6	44.3
Total cholesterol ≥5.18 mmol/l (%)	80.0	61.1	100.0	100.0	69.8	58.8	47.3	38.6
HDL cholesterol <1.04 mmol/l (%)	21.7	16.6	31.7	34.2	13.1	12.3	17.9	21.6
Triglycerides ≥1.70mmol/l (%)	53.2	48.2	100.0	100.0	26.8	26.1	26.7	28.1

Smoking, current-smoker at any examination; hypertension, systolic blood pressure ≥140 mmHg, diastolic blood pressure ≥90 mmHg, use of antihypertensive drugs, or any combination of these at any examination; diabetes, self-reported disease, use of insulin, use of oral hypoglycaemic drugs, non-fasting plasma glucose ≥11.1 mmol/l or any combination of these at any examination. Body mass index was categorized according to World Health Organization definitions. Recommendations of the National Cholesterol Education Program Expert Panel, National Institutes of Health, were used to define dyslipidaemic subgroups (National Cholesterol Education Program, National Heart, Lung, and Blood Institute 2002). To convert total cholesterol and HDL cholesterol to mg/dl, divide by 0.0259; to convert triglycerides to mg/dl, divide by 0.0113. CCHS, Copenhagen City Heart Study; CGPS, Copenhagen General Population Study. HDL, high-density lipoprotein; IHD, ischaemic heart disease.

total cholesterol most likely explained by the introduction of statins in the mid 1990s and onwards.

Our study has strengths and limitations that deserve consideration. The highest risk in both genders in both samples was associated with subgroups including individuals aged ≥65 years with hypertension and/or diabetes, thus ensuring meaningful results generated from this model-building method. In further support of the multi-model strategy, cumulative incidences for IHD generated from the PRIM model were validated for most subgroups in a large independent population. The reason for the validation being divergent for female subgroup 5 could be due to a low number of IHD cases in this subgroup. An additional explanation may be that in this particular subgroup, the fraction of women with total cholesterol ≥5.18 mmol/l (200 mg/dl) was much lower in the CGPS (31.1%) compared with the CCHS (81.6%). The utility of multi-model approaches to improving prediction will depend critically on having model-building and model-validation samples from the same population of inference, as in the present study, because aetiological heterogeneity will be minimal and context will be largely similar. The 94 genetic variants were selected based on previous candidate studies using single-model statistical methods; hence results from these previous studies are not directly comparable to the present findings generated by a multi-model approach. Further, a biological mechanism underlying a potential context-dependent effect of the hepatic lipase (*LIPC*) N193S variant in high-risk subgroup 1 in women remains unclear. Why only one SNV was selected in the PRIM procedure remains unexplained. This may be attributable to the fact that each of the 94 common SNVs were considered to have small to moderate effect sizes compared with the larger effects of each of the traditional risk factors that defined the partitions. Nevertheless, the present strategy serves as an example for future studies of the combined roles of traditional risk factors and the more extensive genomic information that is becoming widely available in population-based studies of the common diseases.

In conclusion, our study has shown that a multi-model strategy has utility for identifying subgroups of individuals characterized by specific contexts with substantial higher risk of IHD than the overall risk for the general population.

Funding

This work was supported by the Danish Medical Research Council [10–081618 to R.F.S.] and the National Institutes of Health [HL072905 and GM065509 to C.S.F.].

Acknowledgements

We thank Kenneth G Weiss, Department of Human Genetics, University of Michigan, for his dedicated attention to the details of programming and analysis of data, and Mette Refstrup, Department of Clinical Biochemistry, Rigshospitalet, Copenhagen University Hospital, for expert technical assistance. Those acknowledged have confirmed their agreement. We are indebted to the staff and participants of the Copenhagen City Heart Study and the Copenhagen General Population Study for their important contributions to our study.

Conflict of interest: None declared.

References

1. Roger VL, Go AS, Lloyd-Jones DM *et al*. Executive summary: Heart disease and stroke statistics – 2012 update. A report from the American Heart Association. *Circulation* 2012;**125**:188–97.
2. Thomas D. Gene-environment-wide association studies: Emerging approaches. *Nat Rev Genet* 2010;**11**:259–72.
3. Sing CF, Stengård JH, Kardia SL. Genes, environment, and cardiovascular disease. *Arterioscler Thromb Vasc Biol* 2003;**23**:1190–96.
4. Bookman EB, McAllister K, Gillanders E *et al*. Gene-environment interplay in common complex diseases: Forging an integrative model - recommendations from an NIH workshop. *Genet Epidemiol* 2011;**35**:217–25.
5. Manolio TA, Collins FS, Cox NJ *et al*. Finding the missing heritability of complex diseases. *Nature* 2009;**461**:747–53.
6. Manolio TA. Genomewide association studies and assessment of the risk of disease. *N Engl J Med* 2010;**363**:166–76.
7. Bodmer W, Bonilla C. Common and rare variants in multifactorial susceptibility to common diseases. *Nat Genet* 2008;**40**:695–701.
8. Teslovich TM, Musunuru K, Smith AV *et al*. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* 2010;**466**:707–13.
9. Holmes MV, Harrison S, Talmud PJ, Hingorani AD, Humphries SE. Utility of genetic determinants of lipids and cardiovascular events in assessing risk. *Nat Rev Cardiol* 2011;**8**:207–21.
10. Gerstein MB,Kundaje A, Hariharan M *et al*. Architecture of the human regulatory network derived from ENCODE data. *Nature* 2012;**489**:91–100.
11. ENCODE project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012;**489**:57–74.
12. Dyson G, Frikke-Schmidt R, Nordestgaard BG, Tybjærg-Hansen A, Sing CF. An application of the Patient Rule Induction Method for evaluating the contribution of the apolipoprotein E and lipoprotein lipase genes to predicting ischemic heart disease. *Genet Epidemiol* 2007;**31**:515–27.
13. Dyson G, Frikke-Schmidt R, Nordestgaard BG, Tybjærg-Hansen A, Sing CF. Modifications to the Patient Rule-Induction Method that utilize non-additive combinations of genetic and environmental effects to define partitions that predict ischemic heart disease. *Genet Epidemiol* 2009;**33**:317–24.
14. Dyson G, Sing CF. Efficient identification of context dependent subgroups of risk from genome-wide association studies. *Stat Appl Genet Mol Biol* 2014;**13**:217–26.
15. Stengård JH, Dyson G, Frikke-Schmidt R, Tybjærg-Hansen A, Nordestgaard BG, Sing CF. Context-dependent associations between variation in risk of ischemic heart disease and variation in the 5'promoter region of the apolipoprotein E gene in Danish women. *Circ Cardiovasc Genet* 2010;**3**:22–30.
16. Frikke-Schmidt R, Nordestgaard BG, Stene MCA *et al*. Association of loss-of-function mutations in the ABCA1 gene with high-density lipoprotein cholesterol levels and risk of ischemic heart disease. *JAMA* 2008;**299**:2524–32.
17. Zacho J, Tybjærg-Hansen A, Jensen JS, Grande P, Sillesen H, Nordestgaard B. Genetically elevated C-reactive protein and ischemic vascular disease. *N Engl J Med* 2008;**359**:1897–908.
18. Kamstrup PR, Tybjærg-Hansen A, Steffensen R, Nordestgaard BG. Genetically elevated lipoprotein(a) and increased risk of myocardial infarction. *JAMA* 2009;**301**:2331–39.
19. Schnohr P, Jensen G, Lange P, Scharling H, Appleyard M. The Copenhagen City Heart Study. Østerbroundersøgelsen. Tables with data from the third examination 1991-1994. *Eur Heart J* 2001;**3**(Suppl H):1–83.
20. Task Force on the Management of Stable Angina Pectoris of the European Society of Cardiology. Guidelines on the management of stable angina pectoris: Executive summary. *Eur Heart J* 2006;**27**:1341–81.
21. Joint European Society of Cardiology / American College of Cardiology Committee. Myocardial infarction redefined -A consensus document of the Joint European Society of Cardiology / American College of Cardiology Committee for the Redefinition of Myocardial Infarction. *Eur Heart J* 2000;**21**:1502–13.
22. The Joint ESC/ACCF/AHA/WHF Task Force for the Redefinition of Myocardial Infarction. Universal definition of myocardial infarction. *Eur Heart J* 2007;**28**:2525–38.
23. Friedman JH, Fisher NI. Bump hunting in high-dimensional data. *Stat Comput* 1999;**9**:123–43.
24. Carlborg O, Haley CS. Epistasis: too often neglected in complex trait studies? *Nat Rev Genet* 2004;**5**:618–25.
25. Cordell HJ. Detecting gene-gene interactions that underlie human diseases. *Nat Rev Genet* 2009;**10**:392–404.
26. Moore JH, Williams SM. Epistasis and its implications for personal genetics. *Am J Hum Genet* 2009;**85**:309–20.
27. Cheverud JM, Routman EJ. Epistasis and its contribution to genetic variance components. *Genetics* 1995;**139**:1455–61.
28. Sanyal A, Lajoie BR, Jain G, Dekker J. The long-range interaction landscape of gene promoters. *Nature* 2012;**489**:109–15.
29. Moore JH, Williams SM. Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. *BioEssays* 2005;**27**:637–46.
30. Wang X, Elston RC, Zhu X. The meaning of interaction. *Hum Hered* 2010;**70**:269–77.
31. Den Dunnen JT, Antonarakis E. Nomenclature for the description of human sequence variations. *Hum Genet* 2001;**109**:121–24.